COMBINING THE DETECTION AND CORRECTION OF SPEECH REPAIRS

Peter A. Heeman,† Kyung-ho Loken-Kim† and James F. Allen†

†University of Rochester
Department of Computer Science
Rochester, NY 14627, USA
{heeman,james}@cs.rochester.edu

‡A TR Interpreting Telecommunications Research Laboratories
22-2 Hikaridai, Seika-cho, Soraku-gun
Kyoto, 619-02, Japan
kyungho@itl.atr.co.jp

ABSTRACT

Previous approaches to detecting and correcting speech repairs have for the most part separated these two problems. In this paper, we present a statistical model of speech repairs that uses information about the possible correction to help decide whether a speech repair actually occurred. By better modeling the interactions between detection and correction, we are able to improve our detection results.

1. INTRODUCTION

Interactive spoken dialog provides many new challenges for spoken language systems. One of the most critical is the prevalence of speech repairs. Speech repairs are dysfluencies where some of the words that the speaker utters need to be removed in order to correctly understand the speaker’s meaning.

Fortunately for the hearer, speech repairs tend to have a standard form. As illustrated in the example below from the TRAINS corpus (d92a-5.2 utt34), they can be divided into three intervals, or stretches of speech: the reparandum, editing terms, and alteration.1

we’ll pick up a tank of uh the tanker of oranges

The reparandum is the stretch of speech that the speaker intends to replace, and this could end with a word fragment, where the speaker interrupts herself during the middle of the current word. The end of the reparandum is called the interruption point and is often accompanied by a disfluency in the intonational contour. This is then optionally followed by editing terms, which can either be a filled pause, such as “um” or “uh” or a cue phrase, such as “I mean”, “well”, or “let’s see”. The last part is the alteration, which is the speech that the speaker intends as the replacement for the reparandum. In order to correct a speech repair, the reparandum and the editing terms need to be identified in order to determine what the speaker intended to say.

1Notation adapted from [12]. Following [14], we use reparandum to refer to the entire interval being replaced, rather than just the non-repeated words. We have made the same change in definition for alteration.

The strategies that a hearer can use for correcting speech repairs depends on the type of repair. For fresh starts, the hearer must determine the beginning of the current utterance and takes this as being the onset of the reparandum. For modification repairs, the hearer can make use of the repair structure, the parallel structure that often exists between the reparandum and alteration, to determine the extent of the reparandum. For abridged repairs, there is no reparandum, and so simply knowing that it is abridged automatically gives the correction.

Previous work in correcting speech repairs [10, 11, 12] has assumed that speech repairs are accompanied by an acoustic editing signal. Given the interruption point, the type of repair, and the syntactic categories of the words involved, Hindle achieved a 97% correction rate and Kikui and Morimoto achieved a 94% correction rate in a Japanese corpus.

However, a reliable acoustic signal has yet to be found [2]. Rather,
The T RAINS corpus consists of 55,000 words of spoken dialogue, where one person pretends to be a computer. Rather, the user knows computer interaction as possible, but was not a scenario.

As part of the T RAINS project [1], which is a long term research project to build a conversationally proficient planning assistant, we have collected a corpus of problem solving dialogs [8]. The dialogs involve two human participants, one who is playing the role of a user and has a certain task to accomplish, and another who is playing the role of the system by acting as a planning assistant. The collection methodology was designed to make the setting as close to human-computer interaction as possible, but was not a wizard scenario, where one person pretends to be a computer. Rather, the user knows that he is talking to another person.

The T RAINS corpus consists of 55,000 words of spoken dialogue, totaling 6 and a half hours of speech. There are 915 modification repairs and 633 fresh starts, not including those whose reparandum is just a word fragment. These repairs have been hand-annotated. There is typically a correspondence between the reparandum and the alteration, and following [2], we annotate this using the labels m for word matching and r for word replacements (words of the same syntactic category). Each pair is given a unique index. Other words in the reparandum and alteration are annotated with an x. Also, editing terms (filled pauses and clue words) are labeled with et, and the interruption point with ip, which will occur before any editing terms associated with the repair, and after the fragment, if present. The interruption point can also be marked as to whether the repair is a fresh start or a modification repair, in which cases, we use ip:can and ip:mod, respectively. The example below illustrates how a repair is annotated in this scheme.

```
engine two from Elmi- or engine three from Elmira
m1 r2 m3 m4 et m1 r2 m3 m4 ip:mod
```

Further details of this annotation scheme can be found in [6].

3. DETECTING SPEECH REPAIRS

For detecting speech repairs, we use a statistical model based on a part-of-speech tagger. Part-of-speech tagging is the process of finding the best category assignment $C_{i,N}$ for a sequence of words $w_{1,N}$. Adopting a probabilistic approach, we want to find the category assignment that is most probable given the words $w_{1,N}$.

$$\max_{C_{i,N}} P(C_{i,N} | w_{1,N})$$

For spoken dialogue, the category transition probabilities from the last word of the reparandum to the first word of the alteration have a different distribution than category transitions for fluent speech. By giving these different distributions to a part-of-speech tagger, the tagger can decide if a transition signals a modification repair, a fresh start, an abridged repair or fluent speech. It can also be used to determine intonational phrase endings and the occurrence of editing terms. For tagging speech repairs, we use the variable $R_i$ to signal the repair transition type between words $w_{i-1}$ and $w_i$. We define $R_i = \text{Mod}$ if $w_i$ is the onset of the alteration of a modification repair. Likewise, we define $R_i = \text{Can}$ for fresh starts and $R_i = \text{Abr}$ for abridged repairs. In the case of no repair, we define $R_i = \text{null}$.

For our probabilistic model, we need to estimate the probability of the repair transition type $R_i$ given the previous context (as well as the probability of $C_i$ given $R_i$ and the previous context, and the probability of $w_i$ given $C_i$; $R_i$ and the previous context). Besides the word categories of the preceding words, this context can also include the presence of editing terms, silence, and word matches [5, 14]. Let $\text{Context}_{t-1}$ be the words preceding $w_{t-1}$, including their category, repair transition, and editing term transition assignment, and let $S_i$ denote the length of silence between word $w_{i-1}$ and $w_i$, and $M_i$ denote the presence of word matches that cross the repair transition. Thus we need to estimate the probability distribution $P(R_i | \text{Context}_{t-1}, S_i, M_i)$. Due to sparseness of data, it is advantageous to make some independence assumptions so that we can model each of these factors independently. If we assume that $\text{Context}_{t-1}$, $S_i$, and $M_i$ are independent, and independent given $R_i$, we can rewrite the probability distribution as the following (through several applications of Bayes Rules).

$$P(R_i | \text{Context}_t) \cdot P(R_i | S_i) / P(R_i) \cdot P(R_i | M_i) / P(R_i)$$

For the probability $P(R_i | M_i)$, we follow [5] and take into consideration the number of intervening words for the closest match and its POS category. In the case where there is more than one word matching at the closest distance (as in the example given in Section 2), we assume

$$P(R_i | \text{Context}_t) = P(R_i | \text{Context}_t) \cdot P(R_i | S_i) / P(R_i) \cdot P(R_i | M_i) / P(R_i)$$

If there are no editing terms for the repair, $R_i = \text{Mod}$ corresponds to the interruption point of the repair. Otherwise, the modification transition type follows the editing terms, which allows the editing terms to be used as part of the context in deciding the type of repair.
use the set of POS categories. To estimate the probability distribution, we use a decision tree algorithm [3] that can ask questions of the following form.

- Is there a word matching across this transition point?
- Is the number of intervening words (between the closest word match) less than \( x \)?
- Does one of the (closest) matching words have a POS category in the subset \( x \) (e.g. is one of the matching words either a common noun or a common plural noun)?

4. CORRECTING SPEECH REPAIRS

After a speech repair has been detected, we need to correct it. In [5], we used well-formedness constraints to determine the correction of modification repairs. The well-formedness constraints make use of word correspondences to find the parallel structure that often exists between the reparandum and the alteration. These word correspondences consist of both word matches and word replacements based on POS tags given by the statistical model. The repair structure is built by using the constraints to limit what can be added to it. Given the interruption point of a speech repair and the POS assignment, the correction algorithm should be able to find the word correspondences, like those given for the example in Section 2. In fact, running the correction algorithm against the correct POS assignments and correct interruption points of modification repairs [9] yields results comparable with those reported by Hindle [10] and Kikui and Morimoto [11].

As for fresh starts, we take the onset of the reparandum as the first word prior to the interruption point that is either the beginning of an intonational phrase, beginning of the turn, or the first word after a single word acknowledgment, such as “Okay”, or “Right” (we use the POS tag AC to denote such words).

5. COMBINING

We already mentioned that the presence of word matches can be used as evidence that a speech repair occurred. However, this only gives a rough indication of whether a repair occurred. What we really want to know is whether the word matches form a suitable repair structure. So, it makes sense to run the correction algorithm on the alternatives being considered by the detection algorithm, and to use the repair structure, along with the word matches that it implicates, as part of the context used for determining whether a speech repair occurred.

Due to the number of possible repair structures and combinations of matching words and the sparseness of data that this would cause, it is not possible to directly use the proposed repair structures in estimating the probability of a repair. Instead, we have extracted a number of features that we feel categorize the proposed repair structures. We then estimate the probability of a repair by feeding these features to a decision tree algorithm, which can decide which features are relevant. We then use this probability estimate in Equation (1) in place of the probability of a repair given the less conclusive evidence of word matches alone.

Below we give the features that the decision tree algorithm uses for estimating the probability of a repair given the proposed repair structure.

Word Matches: As with \( M_i \), we allow the decision tree to ask membership questions about the POS tags of the matching words that have been identified. We also allow it ask about the number of word matches involved in the repair, the number of word replacements, the number of inserted and deleted words, the length of the reparandum, and the length of the alteration.

Amount of Changed Material: For modification repairs, there is typically one sequence of words that have been changed between the reparandum and the alteration, or a sequence of words that have been inserted, and the rest of the alteration simply repeats the reparandum. Consider the following example (d93-14.1 utt10) of fluent speech that has strong parallel correspondences.

it could either take you 8 hours or it could take you 6 hours

\[ m \, m \, x \, m \, m \, r \, m \uparrow \, x \, m \, m \, m \, r \, m \]

Here, the proposed alteration deletes the word “either”, replaces “eight” by “six”, and inserts “or”, giving 3 regions of changed words. In addition to the number of changes between the reparandum and alteration, we also use the size of the largest changed sequence as a feature.

Interruption Point: Another feature is the number of intervening words from the interruption point to the closest word that is marked as a word matching, in both the reparandum and alteration. If the last word of the proposed reparandum has a word matching, then the interruption point is constrained by the reparandum. Likewise if the first word of the alteration has a word matching, then the interruption point is constrained by the alteration. One would expect that the more constrained the interruption point is, the more likely the proposed correction corresponds to an actual speech repair.

Inconsistent Matches: If the proposed interruption point is not constrained by the word matches of the repair structure, it might be the result of the word matches belonging to a repair on a neighboring transition. Consider the potential interruption point identified in the example below (d92a-3.2 utt45).

which engine are we are we taking

\[ m_1 \uparrow \, x \, m_1 \]

The word matching in the proposed repair structure (on the word “are”) in fact belongs to the speech repair whose interruption point occurs on the next transition. The evidence against the proposed interruption point is that not only is the proposed interruption point not constrained by the alteration, but more importantly, there is a word matching, namely the one involving the word “we”, that is consistent with the other word matchings in the repair structure, but does not straddle the proposed interruption point, and so is inconsistent with it. So, we have modified the correction algorithm so that once it has found a repair structure for a proposed interruption point, it checks for the presence of a word matching that is inconsistent
with the proposed interruption point, but for which there is some other transition point for which the combined set of word matchings is consistent.

6. RESULTS

To test out the effect of combining detection and correction, we tested a baseline model in which speech repairs are corrected after the statistical model is run. The baseline model does use the presence of word matchings as evidence that a repair occurred, and then corrects the speech repairs after all repairs have been detected. We also ran the combined model, which uses the proposed repair structure as evidence that a repair occurred, and uses the corrected speech as part of the subsequent context. The results, given in Table 1, were obtained from a 6 fold cross validation test.

In comparison to the baseline results, we find that for modification repairs, detection recall increases from 74.9% to 80.8%, while precision increases from 76.9% to 79.5%. This makes for an improvement in the recall rate of 7.3%, and an improvement in the precision rate of 3.4%. For correcting modification repairs, we also see a similar increase in performance over the baseline model.

For fresh starts, there is also an improvement, but less pronounced than the improvement for the modification repairs. This is because the correction evidence that we are giving the detection routine is tailored for modification repairs, which tend to exhibit the parallel structure between the reparandum and alteration more so than fresh starts do.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>Modification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detection</td>
<td>74.9%</td>
<td>76.9%</td>
</tr>
<tr>
<td>Correction</td>
<td>70.3%</td>
<td>72.2%</td>
</tr>
<tr>
<td>Fresh Starts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detection</td>
<td>57.0%</td>
<td>62.7%</td>
</tr>
<tr>
<td>Correction</td>
<td>46.1%</td>
<td>50.7%</td>
</tr>
</tbody>
</table>

Table 1: Results from treating detection and correction as two separate processes, and from combining them.

7. CONCLUSIONS

In this paper, we illustrated that the two problems of detecting speech repairs and correcting them are not separable. First, the detection algorithm should not only detect the occurrence of a repair, but should also classify the repair based on the correction strategy, be it a fresh start, modification repair, or an abridged repair. Second, the correction strategy should not only propose a correction, but also be able to categorize it in terms of how strongly the repair structure supports the hypothesis that a repair actually occurred. This will help the detection model skip over transitions that should be ruled out by the lack of a convincing repair structure.

8. ACKNOWLEDGMENTS

We wish to thank Laurie Fais, and Tsuyoshi Morimoto. Part of this work was carried out while the first author was visiting ATR Interpreting Telecommunications Research Laboratory. Additional funding was gratefully received from NSF under Grant IRI-90-13160, and from ONR/DARPA under Grant N00014-92-J-1512.

9. REFERENCES